The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Graph Scan Statistics With Uncertainty

Jose Cadena, Arinjoy Basak, Anil Vullikanti
Department of Computer Science and Biocomplexity Institute,
Virginia Tech, Blacksburg, VA 24061
{jcadena,arinjoyb,vsakumar}@vt.edu

Xinwei Deng
Department of Statistics, Virginia Tech, Blacksburg, VA 24061
xdeng@vt.edu

Abstract

Scan statistics is one of the most popular approaches for anomaly detection in spatial and network data. In practice, there are numerous sources of uncertainty in the observed data. However, most prior works have overlooked such uncertainty, which can affect the accuracy and inferences of such methods. In this paper, we develop the first systematic approach to incorporating uncertainty in scan statistics. We study two formulations for robust scan statistics, one based on the sample average approximation and the other using a max-min objective. We show that uncertainty significantly increases the computational complexity of these problems. Rigorous algorithms and efficient heuristics for both formulations are developed with justification of theoretical bounds. We evaluate our proposed methods on synthetic and real datasets, and we observe that our methods give significant improvement in the detection power as well as optimization objective, relative to a baseline.

1 Introduction

Identifying anomalous "hotspots" or outliers is an important problem in the analysis of spatio-temporal and network data, with a large number of applications in areas such as disease surveillance, security, systems biology and social network analysis. One of the popular methods for anomaly detection is the *scan statistics* approach e.g., (Kulldorff 1997; Neill 2012). Informally, this involves formalizing a notion of "anomalousness" by a form of hypothesis testing, based on either an underlying model of the expected data (referred to as *parametric* methods), or based on historical values of the data (i.e., *non-parametric* methods). One of the earliest uses of scan statistics was for finding unusual disease clusters (Kulldorff 1997). See (Akoglu, Tong, and Koutra 2014) for a survey and Section 7 for more details.

Most existing work in anomaly detection for network data typically assumes that the datasets are taken "as is", which is often not a realistic assumption. The observed counts in data have uncertainty and do not exactly match the real world due to reporting errors, geocoding errors, missing entries, etc. All these sources of uncertainty would affect the problem formulations and algorithms for anomaly detection, but they are especially relevant when using scan statistics because the anomaly score is formalized in terms of the log likelihood of occurrence of observed data. One of the very few results on the impact of uncertainty in scan statistics is by (Malizia 2013), who observes that uncertainty exists and that it can affect the quality of the clusters discovered by scan statistics, but this study does not propose any methods for taking uncertainty into account. To the best of our knowledge, the only scan-statistics-based method that incorporates uncertainty is the Bayesian scan statistic proposed in (Neill et al. 2009) and later extended to the multivariate case (Neill and Cooper 2010)—though the authors do not explicitly motivate their work as a solution to uncertainty. Both papers assume simple conjugate priors and thus are able to derive closed-form expressions. However, in general, more complex distributional assumptions for uncertainty would lead to expressions that do not have a closed form and can only be approximately optimized via sampling. In this paper, we use methods from the theory of stochastic optimization to formally characterize scan statistic maximization with uncertainty and develop novel algorithms and heuristics for this problem. Our contributions are summarized below.

- **Scan Statistics with Uncertainty.** We propose two approaches for taking uncertainty into account: a sample average approximation (SAA) and a max-min formulation (Shapiro, Dentcheva, and Ruszczyński 2009; Bertsekas 1995; Prékopa 2013), and show that these are well motivated for scan statistics by analyzing instances with stochastic perturbations. When we account for uncertainty, the anomaly detection task becomes much more challenging. We show that, even without any connectivity constraints, finding clusters based on scan statistics that optimize the max-min objective function is NP-hard, whereas it can be done in linear time if there is no uncertainty.
- **Rigorous Algorithms and Theoretical Bounds.** We develop rigorous algorithms and heuristics for optimizing the scan statistics in both these formulations. For the SAA formulation, our algorithm AGGREGATESAA gives rigorous bounds on the approximation guarantee for finding solutions of a given "effective" size k, with or without connectivity constraints, and is a fixed parameter tractable algorithm. For the max-min formulation, we present two algorithms: (1) MAXMINLPROUND for the case of no connectivity constraints, using linear programming rounding, which yields rigorous approximation bounds, and

(2) a heuristic, BESTMAX, for the case with connectivity. Both AGGREGATESAA and BESTMAX use the technique known as color coding for counting trees, which was adapted for scan statistics by (Cadena, Chen, and Vullikanti 2017).

- **Numerical Evaluation of the Proposed Methods.** We evaluate our proposed methods in two popular benchmarks for scan statistics with known events. Both the SAA and max-min approaches lead to fairly robust scan statistics, and our algorithms have a clear improvement in performance, compared to a natural baseline, over a wide range of signal-to-noise ratio regimes. For regimes with high noise in the data, we see up to two-fold improvement in F1 score. Similarly, the objective score improves in all experimental settings, especially when the signal-to-noise ratio is low, where we see gains over the baseline by more than a factor of 2.

Many details are omitted for brevity, and are available at (Cadena and others 2018).

2 Preliminaries

It is known that scan statistics have been mostly applied to spatial data (Kulldorff 1997). Here, we consider the more general version of scan statistics on networks, which has attracted attention recently (McFowland, Speakman, and Neill 2013; Neill 2012; Speakman and others 2015; Leiserson and others 2015; Hansen and Vandin 2016). Let us define an undirected graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges. For each node $v \in V$, we have an *event* count $c(v)$ and a *baseline* count $b(v)$.

Using scan statistics, the detection of anomalous clusters can be posed as a hypothesis testing problem. The null hypothesis H_0 is that there is no anomalous cluster. That is, the event counts for all nodes are generated independently from the same distribution—proportionally to the baseline counts. Under the alternative hypothesis $H_1(S)$, there exists a small connected subset of nodes S, such that event counts in S are generated at a higher rate than counts elsewhere (i.e., in $V \setminus S$). A scan statistic is an *anomalousness* function $F : S \to \mathbb{R}$ that evaluates how much a subset S deviates from the null hypothesis. Here, we focus on a commonly used scan statistic based on the likelihood ratio test as

$$F(S) = \frac{\Pr(H_1(S)|\text{Data})}{\Pr(H_0|\text{Data})} = \frac{\Pr(H_1(S)|C(S), B(S), C(V \setminus S), B(V \setminus S))}{\Pr(H_0|C(V), B(V))},$$

where $C(S) = \sum_{v \in S} c(v)$ is the total count of S, and $B(S) = \sum_{v \in S} b(v)$ is the baseline count.

Depending on the type of data (i.e., counts, positive real values, etc.) and the statistical assumptions about the process generating the event counts (i.e., Poisson, Normal, etc.), the scan statistics can be broadly summarized in two categories:

Parametric scan statistics. It assumes that observations follow a parameterized distribution, usually from the exponential family, such as Poisson or Normal. For example, the Kulldorff scan statistic for count data, commonly used in disease surveillance (Kulldorff 1997; Duczmal, Kulldorff, and Huang 2006; Kulldorff, Tango, and Park 2003; Neill 2012), is defined as

$$F(S) = C(S)\log\left(\frac{C(S)}{B(S)}\right) + (C(V) - C(S))\log\left(\frac{C(V) - C(S)}{B(V) - B(S)}\right) - C(V)\log\left(\frac{C(V)}{B(V)}\right)$$

if $C(S)/B(S) > C(V)/B(V)$ and 0 otherwise.

Nonparametric scan statistics. The observations are not assumed from a specified distribution. The main idea is to compute the empirical distribution of the counts based on multiple snapshots of the graph. First, one can define a p-value $p(v)$ for each vertex v by comparing the current values of $c(v)$ and $b(v)$ to past observations. Then, a hypothesis test can be performed to check whether the empirical p-values are uniformly distributed on $[0, 1]$ (Qian, Saligrama, and Chen 2014; Sharpnack, Krishnamurthy, and Singh 2013; Neill 2012). For example, in the Berk-Jones scan statistic (BJ) (Berk and Jones 1979), a node is declared to be *significant* if $p(v) < \alpha$ for a given significance level α. The *weight* of a node with respect to α is $w(v, \alpha)$ is 1 if v is significant and 0 otherwise. We use $w(v)$ when α is clear from context. Then, the weight of a set S is $W(S) = \sum_{v \in S} w(v)$. The baseline count is $b(v) = 1$ for all nodes, so $B(S) = |S|$. The BJ scan statistic is defined as

$$F(S) = \max_{\alpha \le \alpha_{max}} |S| \left[\frac{W(S)}{|S|}\log\left(\frac{W(S)/|S|}{\alpha}\right) + \left(1 - \frac{W(S)}{|S|}\right)\log\left(\frac{1 - W(S)/|S|}{1 - \alpha}\right)\right]$$

if $W(S)/|S| > \alpha$ and 0 otherwise.

Thus, the anomaly detection problem in the network can be posed as a constrained optimization problem. Given a graph $G = (V, E)$, a scan statistic $F(\cdot)$, and the associated counts for vertices, C and B, the objective is to find a connected subset $S \subseteq V$, such that $F(S) = F(C(S), B(S))$ is maximized.

Example. Consider an instance of lung cancer cases in a population within a state. We consider a graph $G = (V, E)$ with the set V representing the counties in the state, and E consisting of edges between counties sharing a boundary. The baseline count $b(v)$ is the number of inhabitants in county v, and $c(v)$ is the number of lung cancer cases in v. A cluster of counties in which the incidence counts are significantly different from what is expected based on the population would have a high Kulldorff score.

3 Network Scan Statistics With Uncertainty

The observed data is affected by different sources of uncertainty including unreported cases, inaccurate data collection, measurement error, etc., and how to account for such uncertainty is a crucial question. Let C be a vector of unobserved real counts, such that $c(v)$ is the count associated with node v. Similarly, let X be the vector of observed "noisy" counts. We model uncertainty on X as a distribution θ parameterized by the unobserved real counts C. This distribution could be, for instance, Gaussian noise, such that $X = C + \epsilon$, where ϵ captures our belief about the magnitude of uncertainty. Or a multivariate Gamma distribution with mean C and variance specified by a domain expert. We also define $p(C|\phi)$ as a prior distribution on C controlled by hyperparameters ϕ, which could be inferred from past data or from a domain expert's

knowledge. More generally, let $X = (x(v_1), \ldots, x(v_n))$ and $C = (c(v_1), \ldots, c(v_n))$ be the vectors of observed and real counts, respectively. Then, $X|C \sim \theta(C)$. In most of the existing work, the observed counts are taken as the ground truth; that is, $X = C$ deterministically.

3.1 Problem Formulations

In this work, we consider two approaches to account for uncertainty. The first method is based on the sample average approximation (SAA) method (Shapiro, Dentcheva, and Ruszczyński 2009). The second is based on a max-min formulation (Bertsekas 1995), which is referred as the worst-case scenario in the stochastic optimization (Prékopa 2013).

Scan statistics under sample average approximation. The key idea is taking the expectation of $F(S)$ to account for the uncertainty in scan statistics. That is, given a graph $G = (V, E)$, a scan statistic $F(\cdot)$, and the associated counts for vertices, X and B, the objective is to find a connected subset $S \subseteq V$ that maximizes

$$\mathbb{E}[F(S)|X] = \int_{F(S)} (F(S)|X)p(F(S)|X)dF(S)$$
$$= \int_{C} (F(S)|X)\frac{p(X|C)p(C)}{p(X)}dC.$$

Equivalently, we want to maximize $\int_C (F(S)|X)p(X|C)p(C)dC$, since $P(X)$ is a constant.

The idea of SAA is to generate samples from the distribution for the input and optimize over the set of samples (Swamy and Shmoys 2006). Specifically, maximizing $\mathbb{E}[F(S)|X]$ can be approximated by maximizing the average of N samples in the following manner:

1. for $i = 1$ to N:
   (a) Sample the i^{th} vector of real counts C_i with probability proportional to $p(C_i)$
   (b) Sample the i^{th} vector of estimate counts X_i from $\theta(C_i)$
   (c) Compute $F_i(S) = F(X_i(S), B(S))$ for all connected S
2. Return S^* that maximizes the sample average: $S^* = \arg\max_S \frac{1}{N}\sum_{i=1}^{N} F_i(S)$

Problem 1 *Given a graph $G = (V, E)$, a scan statistic $F(\cdot)$, and N sets of counts for vertices—X_i for $i = 1$ to N—find a connected subset $S \subseteq V$ that maximizes the average score,*

$$\frac{1}{N}\sum_{i=1}^{N} F_i(S) = \frac{1}{N}\sum_{i=1}^{N} F(X_i(S), B(S)).$$

Scan Statistics under the Max-Min Formulation. Rather than taking the average for accounting the uncertainty, an alternative approach to address uncertainty in scan statistics is to consider the minimum score over all the samples. This gives us the following formulation.

Problem 2 *Given a graph $G = (V, E)$, a scan statistic $F(\cdot)$, and N sets of counts for vertices—X_i for $i = 1$ to N—find a connected subset $S \subseteq V$, that maximizes the minimum over all samples,*

$$\min_{i=1,\ldots,N} F_i(S) = \min_{i=1,\ldots,N} F(X_i(S), B(S)).$$

4 Motivation for the two formulations and challenges arising from uncertainty

We use a simple stochastic perturbation model to show that the two formulations, Problem 1 and 2, are well motivated. We consider a simple model using the BJ statistic on an instance $G = (V, E)$ constructed in the following manner. Let V be partitioned into $V = V_1 \cup V_2$, with $w(v) = 1$ for $v \in V_1$, and $w(v') = 0$ for $v' \in V_2$. We also have $|V_1| = k$, so that V_1 is the optimal solution maximizing the BJ statistic for this instance. Next, we assume a simple noise model, in which each non-anomalous node (i.e., those with weight 0) becomes anomalous with a small probability p.

Observation 1 *For the above stochastic model, both the SAA and Max-Min formulations correctly identify the optimal cluster V_1, if the number of samples $N \geq \frac{2\ln n}{\ln 1/p}$. In contrast, the optimal solution for any sample is a subset of V_2, with high probability, missing the real anomalous subgraph V_1.*

Uncertainty makes scan statistics much harder. Neill (Neill 2008) observed that many scan statistics (without any uncertainty), including the BJ-statistic can be solved in linear time if there are no connectivity constraints, because of the "linear time subset scanning" property. As a result, the optimum solution of a given size can be found by considering the items in the order of their counts. In contrast, we show below that under uncertainty, the Max-Min formulation of Problem 2 is NP-complete. We note that maximizing scan statistic score with connectivity constraints was shown to be NP-Hard in (Cadena, Chen, and Vullikanti 2017).

Lemma 1 *For any non-parametric scan statistic $F(\cdot)$, finding a subset $S \subset V$ without any connectivity requirement, and of size at most k, that maximizes $\min_i F_i(S)$ is NP-complete.*

The proof is by a many-to-one reduction from the set multi-cover problem (Chekuri, Clarkson, and Har-Peled 2012), and the main idea is that the optimum solution to the Max-Min objective requires a certain number of anomalous nodes in each sample. By guessing this number, and mapping each set to a sample, the reduction can be made to work. The proof is discussed in the Supplementary Material (Cadena and others 2018).

5 Proposed Methods

We describe our algorithms for the SAA and Max-Min objectives. We focus on non-parametric functions here, though our methods extend to parametric functions in a natural manner. The approximation bounds depend on the specific functions, and here we derive those for the BJ-statistic (see Section 2). For the rest of the section, we use $w_i(v, \alpha)$ to denote the weight of node v—defined in Section 2—in the i^{th} replicate. For brevity, many of the details and proofs are presented in the Supplementary Material.

5.1 Algorithm for the Sample Average Approximation Formulation

Algorithm 1 describes AGGREGATESAA for finding a solution to problem 1. Our method builds on the work of (Cadena,

Algorithm 1: AGGREGATE-SAA$((G(V,E),\alpha_{max}),k,\epsilon)$.

1: **Input**: Instance $(G(V,E),\{\mathbf{w_i}: i=1,\ldots,N\},\alpha_{max})$, parameters k, ϵ
2: **Output**: Set S^* of size at most k
3: Let A be the set of p-values of nodes in V below α_{max}
4: **for** $\alpha \in A$
5: Let $\mathbf{w}$ be a weight vector with $w(v)=\sum_{i=1}^{N} w_i(v,\alpha)$
6: $\{S_j^*(\alpha): j=1,\ldots,k\}=$ MAXWT$(G,\mathbf{w},k,\frac{\epsilon}{n^2})$
7: $S^*=argmax_{j\in K,\alpha\in A}\frac{1}{N}\sum_{i=1}^{N} F_i(S_j^*(\alpha))$
8: **return** S^*
9:
10: **procedure** MAXWT$(G(V,E),\mathbf{w},k,\epsilon')$
11: **Input**: Instance $(G(V,E),\mathbf{w})$ and parameters k,ϵ'
12: **Output**: $\{S_j^*: j\in K\}$, such that S_j^* has weight ψ_j
13: Let $\psi_j=-\infty$ for all $j\in K$
14: **for** $\ell=1$ to $e^k\log(1/\epsilon')$
15: For each node v, pick random color $col(v)\in K$
16: **for** $v\in V, s\in K$
17: $M(v,\{s\})=w(v)$ if $col(v)=s$; $-\infty$ otherwise
18: **for** $v\in V$ and $T\subseteq K$, with $|T|\geq 2$
19: $M(v,T)=\max_{\substack{u\in Nbr(v)\\ T_1,T_2\subseteq T}}\{M(v,T_1)+M(u,T_2)\}$
20: **If** $M(v,T)>\psi_{|T|}$ **update** $\psi_{|T|}=M(v,T)$
21: **return** $\{S_j^*: \sum_{v\in S_j^*} w(v)=\psi_j, \text{for } j\in K\}$

Chen, and Vullikanti 2017), and uses the color-coding technique of Alon et al. (Alon, Yuster, and Zwick 1995). This algorithm finds an **optimal** subgraph of size at most k, where k is a parameter, in time $O(a^k poly(n,m))$, for some constant a—i.e., the running time is polynomial on the size of the graph, but exponential on the solution size. In contrast, a "brute force" approach would need $\binom{n}{k}=O(n^k)$ time to examine every possible connected subset of nodes of size at most k.

Intuition behind the algorithm. The main idea is to color the nodes of the graph using $K=\{1,\ldots,k\}$ colors and restrict the search to "colorful" solutions, which are subgraphs with distinctly colored nodes. This immediately leads to an efficient algorithm because: (1) colorful solutions can be computed using a simple dynamic program, and (2) if the coloring is done randomly, there is a reasonable probability that the optimal solution is colorful. (Cadena, Chen, and Vullikanti 2017) show that the scan statistics problem can be solved by such a dynamic program. Further, combined with a graph compression technique, one can discover solutions with hundreds of vertices while keeping k below 10. The final algorithm is randomized, and it returns an optimal solution with probability $(1-\epsilon)$ in time $O((2\epsilon)^k m\log(1/\epsilon))$, where k is solution size after such a compression. Problems for which such algorithms exist are said to be fixed parameter tractable.

Overview of Algorithm AGGREGATESAA.

- The for loop in lines 4-6 tries out each potential value of α. For each candidate α, the aggregate weight vector $\mathbf{w}$ is computed following the process described in Section 2.
- The subroutine MAXWT (from Cadena, Chen, and Vullikanti) is called in line 6, and it returns a candidate solution $S_j^*(\alpha)$ for each α, and each size $j\leq k$. It uses color coding to find the optimal solution by dynamic programming.

Theorem 1 *Suppose the solution S^* computed by Algorithm* AGGREGATESAA *corresponds to $S_j^*(\alpha)$ for some $j\in K$. Suppose $\sum_{i=1}^{N}\sum_{v\in S_j^*(\alpha)} w_i(v)\geq c\alpha jN$, for a constant $c>1$. Then, the score of S^* is within a factor of $\frac{c\alpha\log 1/\alpha}{c\alpha\log c+(1-c\alpha)\log\frac{1-c\alpha}{1-\alpha}}$ of the optimum score, with probability at least $1-\epsilon$ for any $\epsilon\in(0,1)$. The total running time and space used are $O(2^k e^k|A|Nm\log(n^2/\epsilon))$, and $O(2^k n)$, respectively.*

Proof: (Sketch) The proof relies on the convexity of the function $F(\cdot)$ for non-parametric functions. Since $\sum_{i=1}^{N}\sum_{v\in S_j^*(\alpha)} w_i(v)\geq c\alpha jN$, the minimum value $\sum_i F_i(S_j^*(\alpha))$ can take is when the weight in each replicate is the average, namely $c\alpha j$. Therefore,

$$\sum_i F_i(S_j^*(\alpha))/N\geq c\alpha\log c+(1-c\alpha)\log\frac{1-c\alpha}{1-\alpha}$$

Since $S_j^*(\alpha)$ maximizes the total weight $\sum_{i=1}^{N}\sum_{v\in S_j^*(\alpha)} w_i(v)$, it follows that for any other set S', $\sum_{i=1}^{N}\sum_{v\in S_j^*(\alpha)} w_i(v)\geq\sum_{i=1}^{N}\sum_{v\in S'} w_i(v)$. The maximum value that can be achieved by $\sum_i F_i(S')$ is when the total weight in some replicates is close to j and 0 in the rest (See (Cadena and others 2018)). Therefore,

$$\sum_i F_i(S')/N\leq\frac{\sum_{i=1}^{N}\sum_{v\in S'} w_i(v)}{jN}log1/\alpha=c\alpha\log 1/\alpha$$

The approximation factor is therefore bounded by the ratio of these, which proves the theorem. ■

Performance guarantee in practice. The performance guarantee in Theorem 1 depends on how far the average weight of the sets is from α, over the samples. Empirically, we find that the approximation bound decreases with both c and α, and is very close to 1 in our experiments in Section 6.

5.2 Algorithm for the Max-Min Formulation

The basic idea of our algorithm is to "guess" the total weight z^* of the anomalous nodes in the optimal solution in the minimum sample, and then find a solution that has at least weight z^* in each sample—this corresponds to a multi-cover problem, as in the reduction in the proof of Lemma 1. We use the linear programming rounding method of Kolliopoulos et al. (Kolliopoulos and Young 2005) for finding such an approximate solution. Algorithm 2 describes MAXMINLPROUND for this problem.

Lemma 2 *Let S_{z^*} be the solution returned by Algorithm 2. Then, $\min_i F_i(S_{z^*})\geq\frac{KL(z^*/(k\log n),\alpha)}{KL(z^*(1-\epsilon)/k,\alpha)}$ for any $\epsilon\in(0,1)$.*

Next, we describe the heuristic BESTMAX for problem 2 with connectivity constraints. Proving its approximation guarantee remains an open problem; here, we analyze its running time.

Lemma 3 *Algorithm* BESTMAX *takes time $O(2^k e^k|A|Nm\log(n^2/\epsilon))$ and uses space $O(2^k n)$, where A is the set defined in line 3 of the algorithm.*

Algorithm 2: MAXMINLPROUND$(V, \mathbf{w}, k)$ for Max-Min formulation without connectivity constraints.

Input: Instance $V, \mathbf{w}_i, i = 1, \ldots, N$ and parameter k
Output: Instance $S \subseteq V$ that maximizes $\min_i F_i(S)$ with $|S| = k$
for $z \in [1, \max_i \sum_v w_i(v)]$ in powers of $(1 + \epsilon)$
  Construct matrix $A \in \mathbb{R}^{N \times |V|}$ with $A_{iv} = w_i(v)$ and $b \in \mathbb{R}^N$ with $b_i = z$
  Find a solution $x \in \{0, 1\}^V$ that minimizes $\sum_i x_i$ and satisfies $Ax \geq b$ with the algorithm of (Kolliopoulos and Young 2005)
  if a solution x exists, let $S_z = \{i : x_i = 1\}$
return $S_{z^*} = \{i : x_i = 1\}$ for the maximum z^* for which there is a solution

Algorithm 3: BESTMAX$((G(V, E), \alpha_{max}), k, \epsilon)$ for Max-Min formulation with connectivity constraints.

Input: Instance $(G(V, E), \{\mathbf{w_i} : i = 1, \ldots, N\}, \alpha_{max})$, parameters k, ϵ
Output: Set S^* of size k
Let A be the set of p-values of nodes in V below α_{max}
for $\alpha \in A$
  $\{S^*_j(i, \alpha) : j = 1, \ldots, k\} =$ MAXWT$(G(V, E), \mathbf{w_i}, k, \epsilon/n^2)$
$S^* = argmax_{j \in K, i, \alpha \in A} \min_{i'=1}^{N} F_{i'}(S^*_j(i, \alpha))$
return S^*

6 Experiments

Our experiments are motivated by the following questions:

- **Detection and optimization power.** Does accounting for uncertainty improve detection of anomalous clusters compared to the deterministic case? How do our algorithms AGGREGATESAA and BESTMAX perform for the objective functions in Problems 1 and 2?
- **Effect of the number of replicates** How does detection power vary with the number of replicates N?
- **Approximation guarantee in practice** What are the practical implications of Theorem 1? What is the empirical approximation bound of AGGREGATESAA?

We evaluate our algorithms for scan statistics with uncertainty on the Kulldorff statistic (Kulldorff 1997) and the BJ statistic (Berk and Jones 1979), which are examples of parametric and non-parametric scan statistics, respectively. We evaluate the event detection power in terms of accuracy, precision, recall, and the F1 score. Let R be the set of nodes in the anomalous subgraph we discover and let S be the detected subgraph; then, we define
(1) Accuracy$(R, S) = \frac{|R \cap S|}{|R \cup S|}$,
(2) Precision$(R, S) = \frac{|R \cap S|}{|S|}$,
(3) Recall$(R, S) = \frac{|R \cap S|}{|R|}$, and
(4) F1 score $= 2\left(\frac{\text{Precision}_{(R,S)} \cdot \text{Recall}_{(R,S)}}{\text{Precision}_{(R,S)} + \text{Recall}_{(R,S)}}\right)$. We mainly discuss results for the F1 score below, but plots for the other metrics can be found in (Cadena and others 2018).

Baselines. For the evaluation, we compare our algorithms to the following baseline: Given N count vectors as in Problems 1 and 2, we select one of these vectors uniformly at random and return the connected set S that maximizes the scan statistic in the selected counts. The baseline for Problem 1, referred to as B-SAA, returns the SAA objective value for S. Similarly, the baseline for Problem 2, referred to as B-MAX-MIN, returns the Max-Min objective value for S. These baselines reflect the current practice of ignoring uncertainty.

6.1 Datasets

The Northeastern USA Benchmark (NEast). This dataset (Kulldorff, Tango, and Park 2003) corresponds to occurrences of cancer in a network of 245 counties in the Northeastern part of USA. Under the null model (i.e., no significant cluster), each node v has a count $c(v) \sim Poisson(pb(v))$, where $p = 2.03 \times 10^{-5}$. We simulate anomalous clusters in this network as follows: A cluster consists of a node selected at random and all its neighbors. We generated three such clusters, which can be found in (Cadena and others 2018). These clusters vary in difficulty of detection—easy, medium, and hard. The counts for a node v inside the cluster are sampled from $Poisson(qb(v))$. We perform experiments with values of q of the form $q = \beta p$, where $\beta > 1$ is a parameter that we call *signal strength*. Intuitively, nodes in the anomalous clusters have β times as many expected counts as nodes outside the cluster. In Section 6.2, we discuss the effect of this parameter.

Then, we perturb the counts generated above using Gaussian noise. That is, for each node v, we sample and round down a count $x(v) \sim \mathcal{N}(c(v), \sigma^2)$, where σ^2 is a *noise* parameter. Notice that one could have a different noise parameter for each node, but we only consider uniform noise in our experiments.

Battle of the Water Sensor Networks (BWSN) This dataset (Ostfeld and others 2008) was originally used to evaluate different sensor network designs in terms of early detection of contaminants in a water system. The dataset includes "ground truth" subgraphs representing parts of the network that are contaminated at different points in time—there is one ground truth graph for each snapshot. We control the noise on the sensors with a parameter ϵ. With probability ϵ, the real p-value of a sensor in the network is replaced by a random p-value uniformly sampled from the interval $[0, 1]$. We show results for three clusters; these are typical for other clusters as well.

6.2 Evaluation

Detection and optimization performance. First, we discuss the performance of our methods on the NEast benchmark. We evaluate performance for a wide range of signal (β) and noise (σ^2) parameters (defined in Section 6.1). In Figure 1, we show results for SAA (left), Max-Min (center), and Baseline (right) in the medium-difficulty instance. The heatmaps correspond to different combinations of signal and noise, with darker colors indicating a higher F1 score. We observe that our algorithms for SAA and Max-Min obtain higher scores than the baseline for a larger range of β-σ^2 combinations.

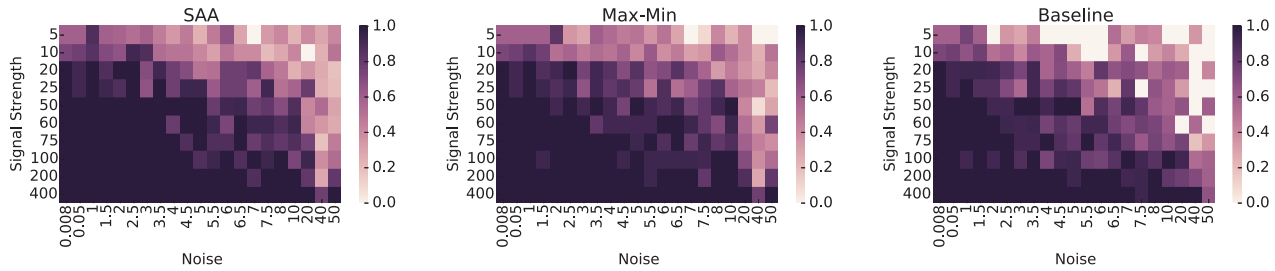

Figure 1: F1 score for various combinations of signal strength and noise for one of the clusters in the NEast dataset, using the AGGREGATESAA and BESTMAX algorithms, and the baseline (darker color means better performance).

In Figure 2, we provide an alternative way to summarize the performance gain for both formulations. The plots show the fraction of signal-noise combinations for which our algorithms have a given percentage of improvement over the baseline. For example, with the SAA formulation (left plot), we obtain at least a 20% improvement (x-axis) over the baseline on 25% (y-axis) of the signal-noise combinations from Figure 1 for the medium-difficulty cluster (green line). The two leftmost plots in Figure 2 reveal that we have larger improvement on F1 score over the baseline on the medium-difficulty instance than in the other two. If a subgraph is easy to discover, accounting for uncertainty may not offer a significant advantage. On the other hand, if an instance is hard to discover, performance is affected in all methods. However, we observe an improvement in performance in all clusters. Finally, here, we can see that the Max-Min formulation has higher improvement for the same instance. The same trends are observed for the objective score (two rightmost plots).

Next, we discuss results for the BJ statistic. In Figure 3 (top), we show the detection performance on the BWSN dataset with the BJ statistic as a function of noise, for three clusters. As expected, performance degrades for all methods as noise increases. However, our algorithms have a better performance than B-SAA and B-MAX-MIN for all levels of noise, and they degrade with the level of noise gradually. In contrast, the baselines show a very inconsistent performance. AGGREGATESAA and BESTMAX also show improved objective score compared to the baseline. In particular, for Problem 2, BESTMAX performs significantly better, typically giving over 20% improvement over B-MAX-MIN.

Effect of N on performance. In Figure 4, we show the F1 score as a function of N, the number of replicates for the BWSN dataset. We observe that the detection power improves as we use more samples for both the SAA and the Max-Min formulation. However, the baseline does not benefit from a larger N.

Approximation bound in practice. We analyze the empirical performance of AGGREGATESAA, compared with the worst case bound derived in Theorem 1. For each snapshot of the BWSN dataset, we compute the approximation ratio given in Theorem 1 for the subgraphs discovered by our algorithm. Figure 5 shows the empirical worst-case guarantee for different noise levels (i.e., each box in the plot) and each snapshot of the dataset (i.e., data points used to draw the box). We see that, for almost all the cases, the approximation guarantee is at least 80%. Further, we observe that the approximation generally becomes worse and has higher variance as noise increases. The increase in the two highest levels of noise may seem counterintuitive, but it is explained by the fact that the bound derived in Theorem 1 is with respect to the size of the solution discovered. As noise increases, the heuristic discovers smaller solutions, which are easier to approximate.

7 Related Work

Our paper is related to the broad area of anomaly detection for network data, and we refer to Akoglu et al. (Akoglu, Tong, and Koutra 2014) for a comprehensive survey on this topic. For brevity, we only discuss work on scan statistics. There are a number of parametric scan statistics, depending on the specific assumption about the observations, e.g., Positive Elevated Mean Scan Statistic (Qian, Saligrama, and Chen 2014), Expectation-based Poisson Scan Statistic (Neill 2012), and Expectation-based Gaussian Scan Statistic (Neill 2012), in addition to the Kulldorff Scan Statistic (Kulldorff 1997) discussed in Section 2. In general, optimizing these functions is challenging in the presence of network constraints, and a number of heuristics have been proposed, e.g., branch-and-bound methods (Speakman and others 2015), Additive GraphScan (Speakman, Zhang, and Neill 2013) based on shortest paths in the graph, semi-definite programming (Qian, Saligrama, and Chen 2014) and Steiner tree heuristics (Rozenshtein and others 2014). The color-coding based algorithm of (Cadena, Chen, and Vullikanti 2017) gives rigorous results for all these functions. Similarly, in addition to the Berk-Jones (Berk and Jones 1979) described in Section 2, there are a number of non-parametric scan statistics, such as Higher Criticism (Donoho and Jin 2004), Kolmogorov-Smirnov (Wilcox 2005) and Anderson-Darling (Eicker 1979). There are several heuristics to optimize such functions, (McFowland, Speakman, and Neill 2013; Neill and Lingwall 2007; Neill 2008; Chen and Neill 2014). The approach of (Cadena, Chen, and Vullikanti 2017) extends to these functions as well.

However, as discussed in Section 1, none of the above methods deal with data uncertainty. Malizia (Malizia 2013) is one of the few papers considering the effect of uncertainty on scan statistics, but not how how to account for it. The only method that we are aware of for incorporating uncertainty is

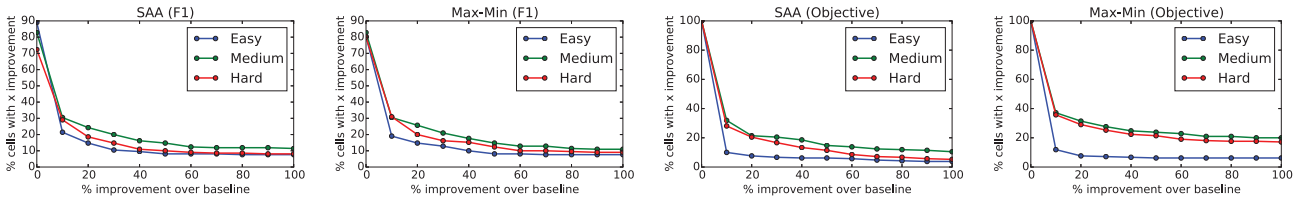

Figure 2: Relative improvement of AGGREGATESAA and BESTMAX over the baseline on the F1 score (left plots) and objective score (right plots) in the NEast dataset. The y-axis corresponds to the fraction of cells in Figure 1, for which our algorithms have a certain level of improvement over the baselines (x-axis). Higher is better.

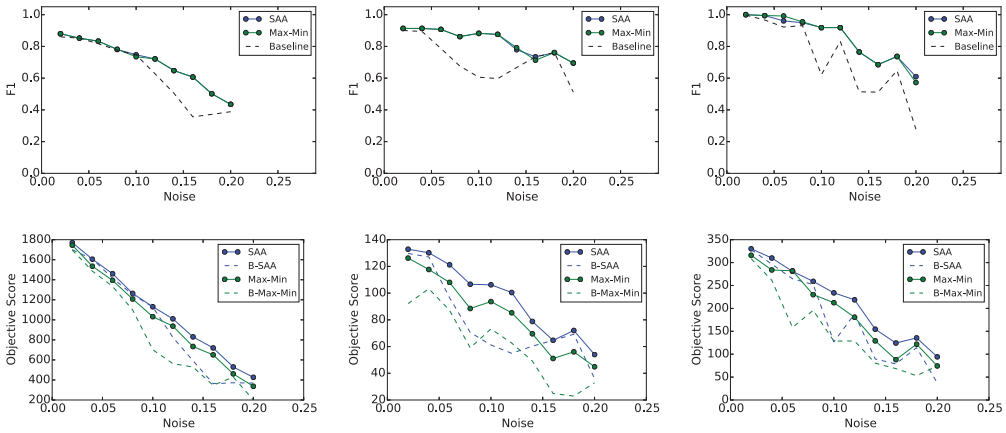

Figure 3: F1 score (top) and objective score (bottom) of AGGREGATESAA and BESTMAX, as a function of the noise level, compared to the baseline for three clusters in the BWSN dataset. Higher is better.

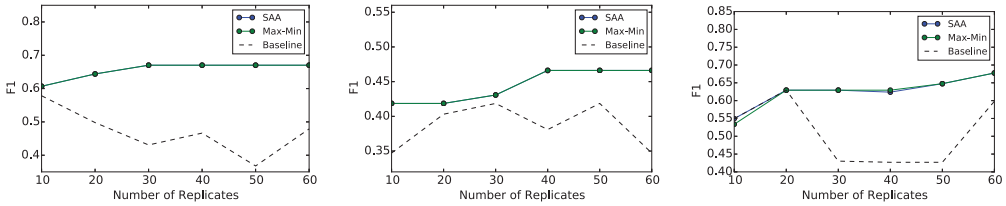

Figure 4: F1 score as a function of N for the BWSN dataset. Detection power increases with the number of replicates for our algorithms, but not for the baseline.

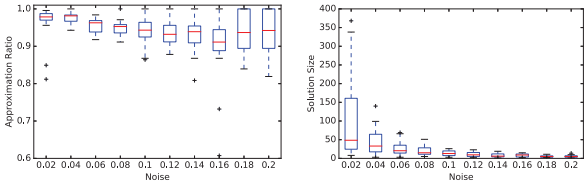

Figure 5: Empirical approximation guarantee of AGGREGATESAA for the BJ statistic (left) and size of the discovered subgraph (right) in the BWSN dataset.

the Bayesian scan statistic proposed in (Neill et al. 2009; Neill and Cooper 2010)—though the authors do not explicitely motivate their work as a solution to uncertainty. The authors propose a Bayesian extension of the Kulldorff statistic assuming a Gamma-Poisson conjugate model, which allows them to derive a closed-form scan statistic. However, the methods that we develop are more general because we don't require a closed-form expression.

8 Conclusions

Scan statistics are used extensively in anomaly detection, but most previous works do not incorporate data uncertainty. We propose the first characterization of the effects of uncertainty on scan statistics using two formulations from stochastic optimization, and we design rigorous algorithms and heuristics for these problems. Our evaluation shows that both ap-

proaches give clear improvement on the detection power relative to a natural baseline. We expect our methodology can help incorporate the effects of uncertainty in other problems as well.

Acknowledgements

This work was partially supported by the following grants: DTRA CNIMS Contracts HDTRA1-11-D-0016-0010, HDTRA1-17-D-0023, and NSF grants IIS-1633028, ACI-1443054.

References

Akoglu, L.; Tong, H.; and Koutra, D. 2014. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*.

Alon, N.; Yuster, R.; and Zwick, U. 1995. Color-coding. *Journal of the ACM (JACM)*.

Berk, R. H., and Jones, D. H. 1979. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete*.

Bertsekas, D. P. 1995. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA.

Cadena, J., et al. 2018. Graph scan statistics with uncertainty. http://tinyurl.com/ybpaa384.

Cadena, J.; Chen, F.; and Vullikanti, A. 2017. Near-optimal and practical algorithms for graph scan statistics. In *SIAM Data Mining (SDM)*.

Chekuri, C.; Clarkson, K. L.; and Har-Peled, S. 2012. On the set multicover problem in geometric settings. *ACM Trans. Algorithms* 9(1):9:1–9:17.

Chen, F., and Neill, D. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *KDD*.

Donoho, D., and Jin, J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*.

Duczmal, L.; Kulldorff, M.; and Huang, L. 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*.

Eicker, F. 1979. The asymptotic distribution of the suprema of the standardized empirical processes. *The Annals of Statistics*.

Hansen, T., and Vandin, F. 2016. Finding mutated subnetworks associated with survival in cancer. *arXiv preprint arXiv:1604.02467*.

Kolliopoulos, S. G., and Young, N. E. 2005. Approximation algorithms for covering/packing integer programs. *Journal of Computer and System Sciences* 495–505.

Kulldorff, M.; Tango, T.; and Park, P. J. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics: Theory and Methods*.

Leiserson, M., et al. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics* 47(2):106–114.

Malizia, N. 2013. Inaccuracy, uncertainty and the space-time permutation scan statistic. *PLoS ONE*.

McFowland, E.; Speakman, S.; and Neill, D. B. 2013. Fast generalized subset scan for anomalous pattern detection. *JMLR* 14(1).

Neill, D. B., and Cooper, G. F. 2010. A multivariate bayesian scan statistic for early event detection and characterization. *Machine learning* 79(3):261–282.

Neill, D. B., and Lingwall, J. 2007. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*.

Neill, D. B.; Cooper, G. F.; Das, K.; Jiang, X.; and Schneider, J. 2009. Bayesian network scan statistics for multivariate pattern detection. In *Scan Statistics*.

Neill, D. B. 2008. Fast and flexible outbreak detection by linear-time subset scanning. *Advances in Disease Surveillance*.

Neill, D. B. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Ostfeld, A., et al. 2008. The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*.

Prékopa, A. 2013. *Stochastic programming*, volume 324. Springer Science & Business Media.

Qian, J.; Saligrama, V.; and Chen, Y. 2014. Connected sub-graph detection. In *AISTATS*.

Rozenshtein, P., et al. 2014. Event detection in activity networks. In *KDD*.

Shapiro, A.; Dentcheva, D.; and Ruszczyński, A. 2009. *Lectures on stochastic programming: modeling and theory*. SIAM.

Sharpnack, J.; Krishnamurthy, A.; and Singh, A. 2013. Near-optimal anomaly detection in graphs using lovasz extended scan statistic. In *NIPS*.

Speakman, S., et al. 2015. Scalable detection of anomalous patterns with connectivity constraints. *Jl Comp Graphical Stat*.

Speakman, S.; Zhang, Y.; and Neill, D. B. 2013. Dynamic pattern detection with temporal consistency and connectivity constraints. In *ICDM*.

Swamy, C., and Shmoys, D. 2006. Algorithms column: Approximation algorithms for 2-stage stochastic optimization problems. *SIGACT News*.

Wilcox, R. 2005. Kolmogorov–smirnov test. *Encyclopedia of biostatistics*.